

Policy Gradient Algorithms Series

ritchie huang

2019-09-18

Contents

1	Preliminaries	2
2	First order approximation on the gradient	2
3	Second order approximation on the constraint	2

1 Preliminaries

From the paper, **TRPO** algorithm finally propose the core optimization problem:

$$\begin{aligned} \theta_{k+1} &= \underset{\theta}{\operatorname{argmax}} \mathcal{L}(\theta_k, \theta) \\ \text{s.t.} \quad & \frac{1}{2}(\theta - \theta_k)^T H(\theta - \theta_k) \leq \delta \end{aligned} \tag{1}$$

2 First order approximation on the gradient

In trust region policy optimization, we finally get this:

$$\mathcal{L}(\theta_k, \theta) = \mathbb{E}_{s, a \in \pi_{\theta_k}} \left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a) \right] \tag{2}$$

if we perform first-order Tylor expansion around the old policy θ_k on this equation, we obtain:

$$\begin{aligned} \mathcal{L}(\theta_k, \theta) &= \mathcal{L}(\theta_k, \theta_k) + \nabla_{\theta} \mathcal{L}(\theta_k, \theta)|_{\theta=\theta_k}(\theta - \theta_k) \\ &= 0 + \nabla_{\theta} \mathcal{L}(\theta_k, \theta)|_{\theta=\theta_k}(\theta - \theta_k) \\ &= \nabla_{\theta} \mathcal{L}(\theta_k, \theta)|_{\theta=\theta_k}(\theta - \theta_k) \end{aligned} \tag{3}$$

we take the gradient w.r.t θ out:

$$\begin{aligned} \nabla_{\theta} \mathcal{L}(\theta_k, \theta)|_{\theta=\theta_k} &= \mathbb{E}_{s, a \in \pi_{\theta_k}} \left[\frac{\nabla_{\theta} \pi_{\theta}(a|s)|_{\theta=\theta_k}}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a) \right] \\ &= \mathbb{E}_{s, a \in \pi_{\theta_k}} [\nabla_{\theta} \log \pi_{\theta}(a|s)|_{\theta=\theta_k} A^{\pi_{\theta_k}}(s, a)] \end{aligned} \tag{4}$$

Coincidentally, the gradient is exactly the gradient of “Vanilla Policy Gradient”.

3 Second order approximation on the constraint